The application of hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification

Post-print of peer-reviewed article published by Elsevier.
Published article available at: http://dx.doi.org/10.1016/j.atmosres.2013.10.019

Cite as:
Malley, C. S., Braban, C. F. and Heal, M. R. (2014) The application of hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification, *Atmospheric Research* 138, 30-40.

The application of hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification

Christopher S. Malley,[1,2] Christine F. Braban[1] and Mathew R. Heal[2]

[1] NERC Centre for Ecology & Hydrology, Bush Estate, Penicuik, EH26 0QB, UK.

[2] School of Chemistry, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JJ, UK.

Corresponding author:

Tel.:+44 7578 725402

E-mail address: C.Malley@sms.ed.ac.uk (C. S. Malley)

Highlights

Classification of 154 EMEP sites between 1991 and 2010 based on ozone variation

Hierarchical cluster analysis reveals four major European ozone regimes.

Non-negative matrix factorization evaluated each site's anthropogenic influence.

The two UK EMEP supersites are representative of the UK's two major ozone regimes.

Auchencorth and Harwell grouped with “Remote” and “Polluted” sites respectively.

Abstract

The effective classification of atmospheric monitoring sites within a network allows conclusions from measurements to be extrapolated beyond the confines of the site itself and applied to larger areas or populations. This is especially important for the European EMEP 'supersites' because these are relatively few in number yet are subject to much investment in composition monitoring capability. Here, the representativeness of the two UK EMEP supersites, Auchencorth and Harwell, was evaluated using hierarchical cluster analysis (HCA) of all available EMEP monitoring sites based on measured ozone concentration datasets for the period 1991-2010. A novel feature was to apply non-negative matrix factorization (NMF) to order the sites within the HCA dendrograms according to the relative anthropogenic influence on ozone. The ordered dendrograms enabled UK sites to be placed more precisely in a European context. For 2007-2010, all 19 UK EMEP sites were assigned to two of the site classification clusters, with 17 of the sites grouping closely with each other in each cluster. Auchencorth clustered with the sites characterised by less modification of hemispheric background ozone levels, while Harwell grouped with the sites showing a more polluted regime. A similar grouping of sites occurred between 1991 and 2010, with relatively closer clustering of Polluted UK sites compared with Remote UK sites due to the larger, transboundary spatial domain for which the Remote UK sites are representative. This tight clustering of the majority of the other UK ozone monitoring sites with either one of the supersites, shows that UK background ozone conditions are well represented by Auchencorth and Harwell, and gives confidence that more extensive chemical climatologies developed for the two supersites will have wider geographical relevance.

Key words: ozone; EMEP monitoring sites; cluster analysis; non-negative matrix factorisation;

1. Introduction

The European Monitoring and Evaluation Programme (EMEP, www.emep.int) provides governments with scientific information to inform policy regarding the long-range, transboundary transport of air pollution (Torseth et al., 2012). The programme has three core strands: collation of atmospheric emissions inventories; modelling of atmospheric transport and deposition; and measurement of atmospheric composition at locations where the impact of local pollutant emission sources should be low. The EMEP guidance (EMEP/CCC-Report 1/95) outlines methods intended to ensure that air sampled at a monitoring site is representative of air not directly affected by local emission sources. These include: 50 km from major pollution sources (towns, power plants etc.), 2 km from the application of manure, and consideration of meteorological and topographical features. EMEP Level I sites are designed to capture basic atmospheric composition, whilst Level II and III sites (often referred to as EMEP supersites) measure a wider range of atmospheric constituents at higher time resolution than Level I (see Torseth et al. (2012)).

Monitoring sites in a network are usually classified into different groups that internally share similar chemical climatologies, i.e. similar atmospheric composition, drivers of that composition, and impacts due to that composition. A balance is required which captures the major variations in composition and drivers across the network but in as few groups as possible so as to retain the ability to generalise. Various grouping methodologies have been applied (Joly and Peuch, 2012). These range from the relatively subjective use of metadata (Spangl et al., 2007), traditionally used in monitoring networks, to more objective techniques

such as rankings based on statistical indicators (Kovac-Andric et al., 2010), linear discriminant analysis (Joly and Peuch, 2012), principal component analysis (Lau et al., 2009), and non-hierarchical (Ignaccolo et al., 2008) and hierarchical cluster analysis (Flemming et al., 2005; Henne et al., 2010; Tarasova et al., 2007). The latter is a multivariate approach that encompasses many separation/agglomeration techniques which aims to identify natural groupings, or clusters, amongst objects in a dataset through minimisation of the within-cluster variance and maximisation of the between-cluster variance (Kaufman and Rousseeuw, 1990). Clustering methods require user-defined parameters which may impact the objectivity of the analysis. For example, a method for calculating 'distance' between individual members needs to be specified (Dabboor et al., 2013), as must a method for calculating the separation between different groups of members (Mangiameli et al., 1996). Nevertheless, as cited above, clustering techniques have been widely applied to grouping atmospheric monitoring sites.

The aim of this study was to assess whether the location of the two Level II UK supersites, at Auchencorth in south-east Scotland and Harwell in southern England, are representative of UK background conditions, even though they do not fully meet the EMEP criteria for non-locally influenced "background" sites (this is the case at other EMEP sites, as acknowledged in Torseth et al. (2012)): Auchencorth is located 17 km from Edinburgh, although prevailing winds mean it is predominantly upwind from the city, whilst Harwell is 7 km from a 1,360 MW natural gas power station. Effective site classification is particularly important for EMEP Level II 'supersites' because these are considerably few in number yet subject to much investment in composition monitoring capability.

In this work, sites across the EMEP domain were classified according to the annual and daily patterns in ground-level ozone concentrations. Ozone was chosen for two reasons. First, it is

the most widely measured constituent across the EMEP network – between 2007 and 2010, 113 sites measured hourly ozone concentrations and 49 sites have continuous ozone time series since 1991. Second, measured ozone concentrations are a result of the combination of a wide variety of drivers which are also relevant to many aspects of atmospheric composition, including precursor emissions, photochemistry, deposition, meteorological and climatic conditions and long-range transport (AQEG, 2009; Royal Society, 2008). A major driver of temporal ozone variation is hemispheric background concentrations (AQEG, 2009). Regional and local-scale processes lead to modification of these values. Under suitable conditions, efficient photochemical processing of NO_x and volatile organic compounds (VOCs) lead to additional ozone formation and high ozone episodes, while local-scale depletion of ozone occurs due to reaction with NO, an effect which increases with higher NO concentrations (Jenkin, 2008).

Hierarchical clustering was applied to the monthly-diurnal ozone concentrations (average diurnal cycle for each month of the year) at each EMEP site over 4-year periods. Although hierarchical clustering has been applied previously to monitoring site classification (Tarasova et al., 2007), the novelty here was the subsequent application of non-negative matrix factorisation (NMF) (Lee and Seung, 2001) to order the sites across the dendrogram according to an extracted factor. In this case the factor represented the extent of anthropogenic influence on ozone concentrations. Hierarchical cluster analysis was chosen in preference to non-hierarchical techniques as the robustness and suitability of the cluster assignment is more objectively investigated through the resulting dendrogram, particularly when this is combined with NMF. By using NMF, the ozone concentrations at the two UK EMEP supersites were placed more precisely in the European context. The analysis was

carried out separately for five 4-year periods spanning 1991-2010 to assess the consistency of site representativeness over time**.**

2. Methodology

Data arrays of 4-year averaged monthly-diurnal ozone concentrations were calculated for each EMEP site, i.e. 288 (= 24 hours × 12 months) ozone concentrations per site where, for example, the ozone concentration for 'Jan-00.00' was the average of the 00.00-01.00 hourly ozone on all days in January in the 4-year period under consideration (1991-1994, 1995-1998, 1999-2002, 2003-2006 and 2007-2010) (Measured data from http://ebas.nilu.no). Four-year averages of monthly-diurnal concentrations were considered a reasonable compromise between long enough to smooth out inter-annual variability whilst short enough to avoid incorporation of long term trends. The number of sites included in each time period, and the number of countries within which these sites are located is summarised in Table 1. 154 sites contributed ozone data to at least one 4-year period, of which 49 contributed to every 4-year time period.

The choice of clustering parameters can impact the clustering result. In this study, the standard Euclidean distance between two *n*-dimensional data arrays was used and in this case $n = 288$ (Equation 1).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \text{Equation 1}$$

In hierarchical clustering each object (here each site's monthly-diurnal ozone concentrations) initially constitutes its own cluster. The two nearest clusters are then combined and this process is continued until there is one cluster containing all objects. The process of agglomeration is summarised in a dendrogram. Hierarchical clustering techniques differ in

how the separation of clusters is quantified, and hence how the linkages of the dendrogram are constructed. A number of linkage methods can be applied, e.g. single, complete, average or centroid linkage, or Ward's method (Kaufman and Rousseeuw, 1990). Mangiameli et al. (1996) applied these linkage methods to model datasets of known cluster assignments to assess their effectiveness under a range of conditions, e.g. with dispersed datasets and large disparities in cluster density. In general, Ward's method resulted in a higher percentage of objects assigned to their correct cluster. Ward's method has also been used previously with pollutant concentration data (e.g. Dillner et al. (2005) and Lu et al. (2006)) and was chosen here. At each step, Ward's method calculates the within-cluster sum of squares (WCSS) for every cluster (Kaufman and Rousseeuw, 1990; Ward, 1963), where WCSS is the squared Euclidean distance between an object in the cluster (x_j) and the mean of that cluster ($\bar{x}$), summed over all (m) objects in that cluster (Equation 2).

$$WCSS = \sum_{j=1}^{m}(x_j - \bar{x})^2 \qquad \text{Equation 2}$$

The two clusters, merged at that step, are those which after merging produce the smallest increase in the sum of WCSS over all clusters.

The branches of the dendrogram are rotatable, allowing the members to be weighted and reordered according to those weightings within the confines of the dendrogram branches. Here, non-negative matrix factorisation (NMF) was used to reorder the monitoring sites based on the range of monthly-diurnal ozone profiles observed across Europe. The 288 ozone concentrations for all monitoring sites are combined into a $n \times m$ matrix, **V**, where n is the number of dimensions (288), and m is the number of monitoring sites (113 for 2007-2010). NMF decomposes **V** into two output matrices, an $n \times r$ matrix, **W**, and an $r \times m$ matrix, **H**, whose product **WH** approximates the input matrix **V** (Figure 1). This approximation is achieved such that the Euclidean distance between the input matrix and output matrix product

(i.e. $(\mathbf{V} - \mathbf{WH})^2$) is minimized. Variable *r* is the number of factors with which to simplify **V**, **H** contains the contribution of each factor at each monitoring site, and **W** details the composition of each factor (Lee and Seung, 1999). Although a locally minimised Euclidean distance between **V** and **WH** resulted from the NMF algorithm, and therefore **W** and **H** varied between runs, this did not cause significant variation in dendrogram reordering results (Lee and Seung, 2001).

Two factors were used in the NMF here; hence **W** described two ozone monthly-diurnal cycles (visualised in Figure 2), from which the monthly-diurnal cycle at each site is reconstructed by considering the contribution of each from **H**. The appearance of the monthly-diurnal cycle of Factor 1 is consistent with an air mass significantly influenced by anthropogenic emissions of ozone precursor/depleting species. It features pronounced diurnal (max ~40 µg m^{-3}) and annual (max ~40 µg m^{-3}) ozone variation, and a summer maximum in ozone concentration, consistent with regional-scale ozone production, as shown in Jenkin (2008). Factor 2 exhibits greater uniformity in ozone concentrations. The choice of two factors was semi-arbitrary, but allows an estimation of the anthropogenic influence at each site to be encapsulated via the contribution from a single factor (Factor 1). Each site was weighted with this contribution, and at each node in the dendrogram, where two branches meet, the mean weight of sites on each branch was calculated. The branch with the higher value was placed on the right hand side of the node, producing a dendrogram reordered based on each site's relative pollution levels.

The hierarchical cluster analysis, NMF and map plotting were undertaken with the R statistical software (R Core Development Team, 2008), using respectively the 'NMFN' (Liu, 2012), 'cluster' (Maechler et al., 2013) and 'openair' (Carslaw and Ropkins, 2013) packages.

3. Results

Figure 3 shows the dendrogram for the EMEP sites recording ozone data for the period 2007-2010. The progression of the average monthly-diurnal ozone cycles across the dendrogram is shown across the bottom of the diagram for selected sites, and illustrates the dramatic change in characteristics of these cycles. Figure 4 shows the proportion of explained within-cluster variance as a function of the number of clusters, and ranges from 100% when all sites were located in individual clusters to zero when all sites were located in one cluster. In selecting the optimal number of clusters, the aim is to maximise the explained inter-site ozone variability using a small number of clusters, allowing key features of ozone variation across Europe to be summarised. The statistics in Figure 4 clearly show that decreasing from 113 to 4 clusters yields relatively small decreases in explained variance (four clusters still explains 75% of within cluster variance) but that a further decrease in cluster number results in substantial disbenefit to explained variance. The natural break into four clusters is also clearly evident in the dendrogram of Figure 3. These four clusters were designated as Remote, Polluted, Elevated and Mountain going from left to right across the dendrogram. The average monthly-diurnal cycles and locations of sites in each cluster are shown in Figures 5 and 6 respectively.

The Remote cluster's average monthly-diurnal cycle comprised an annual maximum in April (~85 $\mu g\ m^{-3}$) and a diurnal maximum during the early afternoon (Figure 5). However, diurnal and annual variations in ozone concentrations were not large. The minimum ozone concentration was ~44 $\mu g\ m^{-3}$ and maximum amplitudes in diurnal and annual ozone variation were 23 and 37 $\mu g\ m^{-3}$ respectively. The majority of the sites in the Remote cluster

are on the north and west fringes of Europe, predominantly in Scandinavia and the UK (Figure 6). In comparison, the Polluted cluster contained significantly more ozone concentration variation. Maximum average ozone concentrations (96 µg m^{-3}) occurred in April and elevated afternoon concentrations persisted through summer before decreasing in August and September. Maximum amplitudes in diurnal and annual ozone variation of 49 and 59 µg m^{-3}, respectively, produced the lowest concentrations across all clusters (minimum concentration: 27 µg m^{-3}). Sites contributing to this cluster were predominantly in central and southern Europe, including central and eastern England (Figure 6). The majority of sites in the Elevated and Mountain clusters are located at altitude. The Mountain cluster contained a greater proportion of mountain-top sites, as opposed to the mix of mountain-top, mountain-side and valley sites found in the Elevated cluster. Diurnal ozone variation was considerably lower in these two clusters than in the Remote and Polluted clusters (maximum amplitude of Mountain cluster diurnal cycle = 5 µg m^{-3}), and the Mountain cluster showed highest ozone concentrations (maximum 119 µg m^{-3}). The average ozone monthly-diurnal cycle of the Elevated cluster had similar features to that of the Mountain cluster, but with lower concentrations and greater annual/diurnal ozone variation. These statistics summarise variation for the period 2007-2010; however, due to numerous factors, including long-term trends and inter-annual variability, these values vary for the other four time periods spanning 1990-2006.

Though the four major clusters explained 75% of the variability in monthly-diurnal ozone variation between EMEP sites for 2007-2010, application of NMF dendrogram reordering allowed the remaining 25% to be investigated. From left to right in Figure 3 sites within each cluster are ordered according to increasing anthropogenic influence. All 19 UK EMEP sites (which are highlighted in Figure 3 and shown geographically in Figure 7) were apportioned

to the Remote and Polluted clusters. Of these, 17 sites grouped closely with each other within the two clusters (Figure 3). The two exceptions were Lerwick and Weybourne. Lerwick, on the Shetland Islands, is much further north and double the distance from a city than any other UK site and was ordered in the dendrogram as considerably more remote than other UK sites in the Remote cluster, i.e. significantly less anthropogenically influenced. Weybourne, on England's east coast, was clustered midway between the two groupings of UK sites, suggesting the site experienced both types of ozone regimes. The Auchencorth EMEP supersite was in the middle of the grouping of twelve UK sites within the Remote cluster, whilst the Harwell supersite grouped with the five UK sites in the Polluted cluster, and was the least anthropogenically influenced of this grouping (Figure 3). It is a highly relevant result that the cluster analysis showed two dominant ozone regimes in the UK and each was well represented by one of the two UK EMEP supersites.

For the other four 4-year periods spanning 1991-2006, sites were included in the clustering if monitor data was available for the 4-year period under consideration. The number of clusters produced for each 4-year time period were not fully consistent. The periods 2007-2010 and 1999-2002 yielded a Remote, Polluted, Elevated, Mountain cluster set. For 2003-2006, the nine least anthropogenically-influenced Remote cluster sites formed an additional, distinct cluster. For 1995-1998, four clusters were produced, but three of these groups were of non-elevated sites, with Mountain and Elevated sites combined into one cluster. For 1991-1994 there were only three clusters: a Remote, Polluted and combined Mountain and Elevated cluster.

The Harwell EMEP supersite was consistently amongst the least anthropogenically-influenced sites of the Polluted cluster in every 4-year period except 2003-2006 (see Figure 3

for 2007-2010). For 2003-2006, the clustering assigned Harwell to the Remote cluster. This shift in classification also occurred for the four other UK sites which clustered tightly with Harwell in 2007-2010 (Market Harborough, Sibton, St Osyth and Wicken Fen). This indicated a change in this UK ozone regime, relative to ozone variation across Europe. This regime, for which Harwell was representative, was characterised by greater diurnal and annual ozone variation than the rest of the UK, and its spatial domain encompassed sites within 120 km of London. In a European context, however, this regime was not the most anthropogenically influenced.

Data submission to EMEP for Auchencorth commenced in 2006, so ozone concentrations from a neighbouring site at Bush, only 8 km from Auchencorth, were used as a proxy for Auchencorth prior to 2006. While small changes to the state and drivers of atmospheric composition can result in relatively large changes in ozone variation, both sites are south of Edinburgh and similarly influenced by local and regional-scale meteorology. In 2011, hourly wind direction between the two sites differed by more than 45^{o} only 3% of the time. For the period 2007-2010, Auchencorth and Bush grouped immediately adjacent in the dendrogram (Figure 3), and the mean difference in hourly ozone concentrations was 6.48 $\mu g\ m^{-3}$. Bush maintained a consistent assignment across the 20-year period, grouping immediately after sites including Zeppelin (on the Arctic island of Svalbard), Mace Head (on the west coast of Ireland), and other more Remote sites. However, the other Remote UK sites exhibited greater variability in their position relative to Bush with time than was the case for the other UK Polluted sites relative to Harwell. For example, in 2003-2006, five UK sites grouped with Bush, however, four were less tightly grouped but remained assigned to the Remote cluster, and one site, (Lullington Heath) was assigned to the Polluted cluster. Similar variability was

found for the remaining time periods. The Remote UK sites were located in the north and west of the UK, with the exception of Lullington Heath on England's south coast.

4. Discussion

Sites within the Remote cluster are further from the largest sources of ozone precursor (NO_x) and depleting (NO) species as shown by their locations predominantly on the north west fringe of Europe. This leads to lower perturbation of larger scale, continental and hemispheric background ozone levels, and hence relatively low amplitude monthly-diurnal ozone cycles. Ozone variation in the Polluted cluster show greater diurnal and annual variation, and sites are closer to major sources of ozone precursor/depleting species which facilitate greater photochemical production during the day and removal at night through deposition and reaction with NO. In the Elevated and Mountain clusters, higher concentrations and less diurnal variability are observed due to ozone transport in the free troposphere. Ozone formed in polluted areas with high concentrations of VOCs and NO_x ventilates from the boundary layer to the free troposphere where lower temperatures reduce ozone loss through reaction with species such as NO (Guo et al., 2013). The greater atmospheric motions at altitude also rapidly replenish any ozone loss by deposition. The lower average altitude of sites in the Elevated cluster possibly leads to a greater degree of local ozone production/depletion, increasing diurnal variability compared with the Mountain cluster.

These results reflect previous analysis of ozone spatial trends across the UK and Europe. Tarasova et al. (2007) used cluster analysis to group 114 extra-tropical sites globally based on ozone variation between 1990 and 2004. Of the six clusters derived, European sites were assigned to five, with only a 'polar/remote' classification not found in Europe. The additional

cluster in Tarasova et al. (2007) compared to this study was because sites assigned to the Remote cluster (for 2007-2010) were separated into Clean Background and Rural clusters. Inspection of Figures 3 and 4 show that this subsidiary separation is not borne out by the application of the clustering algorithm applied here. Ozone variation at monitoring sites in North America, Japan and Argentina was classified in the five clusters containing European sites. Henne et al. (2010) reported an 'observation-independent' Ward's method hierarchical cluster analysis to differentiate 34 rural European sites. Parameters characterising emissions, deposition and transport were included and six clusters ranging from 'generally remote' to 'agglomeration' were identified. The Harwell site grouped in the most polluted 'agglomeration' cluster along with Weybourne. The clustering result explained 55% of the inter-site variability in daily median ozone concentrations, compared with 75% of inter-site variability explained by the methodology adopted here. Comparison of the results from Henne et al. (2010) with the results obtained in this study show one main difference. In the former clustering there was no differentiation between elevated sites (e.g. Jungfraujoch, Switzerland, 3578 m) and non-elevated remote sites. A study of 97 'non-urban' US sites used principal component analysis to identify 14 groups of sites with similar summer ozone variation (Chan and Vet, 2010). Comparison of the monthly ozone variation revealed a spectrum of sites similar to that found in the cluster analysis of European sites, from those with high summer concentrations and a strong influence of regional photochemistry, to those without such regional photochemical influences.

Variation in the clusters produced for each four-year period may be a result of numerous factors. One is long-term trends in ozone variation from the curtailment of ozone precursor emissions, as recently highlighted across 158 EMEP sites by Wilson et al. (2012), and globally by Parrish et al. (2013) who found long-term shifts in seasonal ozone cycles at sites

in Europe and North America. Another factor is changes in the number and distribution of EMEP monitoring sites (Table 1). In 1991-1994, there was only one site in southern Europe (Portugal) and in eastern Europe (Czech Republic). The rest were located in central Europe, Scandinavia and the UK. Hence there was less variability in monthly-diurnal ozone variation overall and only three major clusters formed. By 2010, more sites had been established in Mediterranean regions and eastern Europe providing data across a more varied ozone landscape. A third factor is anomalous characteristics for some time periods. In one 4-year period, 2003-2006, half of the years (2003 and 2006), contained periods when 'heat-wave' conditions affected significant areas across Europe. Lee et al. (2006) detail the significant increase in photochemical activity during these conditions, including an approximate doubling of VOC reactivity with the OH radical. Hence this change in drivers provides another potential method for changes in site classification. In the cluster results from 2003-2006, nine Remote sites formed an additional cluster, suggesting ozone variation at these sites was more different from the other Remote sites than at any other time between 1991 and 2010. Similarly, Polluted UK sites, including Harwell, grouped in the Remote cluster, and therefore showed most difference in ozone variation compared to Polluted sites in central and southern Europe during this period. Conversely, Lullington Heath, on England's south coast, clustered in the Polluted cluster. Its closer proximity to mainland Europe, relative to other UK sites may have led to a greater exposure to the anomalous ozone variation during the two 'heat-wave' periods in 2003 and 2006.

AQEG (2009) reported annual mean ozone concentrations were highest in the north and west UK, but the propensity for prolonged exceedance of health-based ozone metrics was higher in the south and east. This is explained by a greater influence of hemispheric background levels in the UK towards the north-west (Jenkin, 2008), and leads to the separation of UK EMEP

sites found in this analysis. Modification of hemispheric background ozone concentrations occurs through formation of additional ozone on a regional scale due to reaction of NO_x and VOCs, and local-scale depletion of ozone by reaction with NO. Remote UK sites, generally towards the north and west UK are less susceptible to transport of primary emissions from major pollution sources such as southern England and continental Europe (RoTAP, 2012). Hence lower NO concentrations at Remote UK sites lead to less ozone depletion, and greater influence of hemispheric background concentrations. Ozone concentrations at many EMEP sites across Europe, such as in Scandinavia, are similarly influenced by hemispheric background concentrations. The geographic domain for which monthly-diurnal ozone variation at Auchencorth is representative is therefore large and transboundary in nature. Non-UK sites falling within this international domain also cluster tightly with Auchencorth, leading to greater variability in the position of UK Remote sites relative to Auchencorth within the ordered dendrograms.

Conversely, Polluted UK sites, in south-east England, are closer to major sources of VOCs and NO_x and have higher diurnal and annual variability due to greater modification of hemispheric background ozone concentrations. Higher NO concentrations facilitate greater ozone depletion, while the proximity of ozone precursor sources increase the prevalence of regional-scale, elevated ozone episodes. Ozone variation within the entire Polluted cluster is less homogeneous than in the Remote cluster due to differences in emission patterns and other drivers (e.g. solar intensity) between sites. For example, the southern UK has a variety of different drivers determining ozone concentrations compared with sites in central Europe. Depending on meteorological conditions air masses entering this region can contain relatively high or low concentrations of ozone formation and loss species (see e.g. (Jenkin, 2008)). Hence these UK sites, in close proximity to London, cluster more tightly with Harwell

between 1991 and 2010. Harwell is therefore representative of a smaller geographic area than Auchencorth. UK EMEP sites provide an excellent case study for this effect, due to their large number (19) and density. However it is also apparent elsewhere: for example, six sites within 120 km of Vienna all cluster tightly together in the 2007-2010 dendrogram, along with Topolniky, Slovakia, a similar distance from Vienna. These sites (Heidenreichstein, Pillersdorf, Ganserdorf, Stixneusiedl, Illmitz, Dunkelsteinerwald and Topolniky) are located at the more anthropogenically influenced end of the Polluted cluster.

Europe-wide, the clustering produced a number of anomalous classifications. For example, Finokalia, a coastal site on Crete, grouped in the Mountain cluster (Figure 3). It has been shown that significant incursion and entrainment of the free troposphere into the boundary layer occurs at Finokalia producing a monthly-diurnal ozone profile typically found at sites at much higher elevations (Gerasopoulos et al., 2005; Gerasopoulos et al., 2006). Other coastal sites, such as Westerland, Denmark and Cabo de Creus, Spain, were grouped in the Elevated cluster. The NMF reordering facilitated evaluation of other anomalous sites. For example, Lazaropole in Macedonia was a member of the mountain cluster, but the classification was relatively weak and it could be construed as its own cluster. Ozone concentrations at Lazaropole were significantly higher than the 112 other sites in the analysis, with 51% of the monthly-hourly averaged concentrations exceeding 120 $\mu g\ m^{-3}$. The source of this anomaly remains unexplained. There are no other EMEP sites within 300 km of Lazaropole, and only two sites within 500 km, so to examine the geographic extent of the high ozone concentrations reported at Lazaropole, and the ozone regime across south eastern Europe in general, requires a greater number of sites.

Few pollutants are monitored as widely as ozone, with some not routinely measured at all, making pollutant-specific site representativeness studies impractical. For example, peroxyacetyl nitrates (PAN), which provide a mechanism for NO_x storage and long range transport, are not continuously monitored at EMEP sites (McFadyen and Cape, 2005), but it is nevertheless important to assess the spatial relevance of conclusions of sporadic PAN measurement. Since ozone is influenced by a wide variety of drivers, site classifications based on ozone may be anticipated to be representative for many other atmospheric species - particularly secondary components, including PAN. Hence, as a consequence of demonstrating the representativeness of the UK EMEP supersites based on ozone variation, comprehensive chemical climatologies derived for the impacts of other atmospheric components at the sites can be placed in a European context. Application of this methodology to different pollutant datasets could be used to assess changes in site classification as a function of pollutant. For example, a recent study applies a k-means clustering algorithm to differentiate US sites based on five-year averaged concentrations of $PM_{2.5}$ components (Austin et al., 2013).

5. Conclusions

Two major ground-level ozone regimes over the UK have been identified through the use of hierarchical cluster analysis of monthly-diurnal ozone datasets from 154 EMEP monitoring sites across Europe. The application of non-negative matrix factorization (NMF) reordered the summary dendrogram based on the relative anthropogenic influence on ozone at each site. This allows the 25% within-cluster variability in ozone concentrations across Europe not explained by the identification of four major clusters to be interpreted. For 2007-2010, all 19 UK EMEP sites were assigned to two of the Europe-wide clusters, with 17 sites apportioned

into 2 groups in these two clusters. One cluster is comparatively less anthropogenically influenced, with ozone concentrations featuring less modification from hemispheric background levels; the other cluster is of sites closer to ozone precursor/depleting emissions and features more pronounced diurnal and annual ozone cycles. The UK EMEP supersites of Auchencorth and Harwell grouped tightly with the other UK sites in these 'Remote' and 'Polluted' clusters respectively. For the other four, four-year periods considered between 1991 and 2006, a similar separation of UK sites occurred, with relatively tighter clustering of Polluted UK sites to Harwell than Remote UK sites to Auchencorth/Bush due to the larger, transboundary spatial domain for which Auchencorth is representative. Hence the UK background ozone conditions are well represented by the location of the UK EMEP supersites at Auchencorth and Harwell. Both supersites currently monitor 120 different chemicals in air, precipitation and particulate matter; this work shows that conclusions derived from interpretation of these large datasets are likely appropriately applied to a wider geographic area.

Acknowledgements

C. S. Malley acknowledges the University of Edinburgh School of Chemistry, the NERC Centre for Ecology & Hydrology (CEH) and the UK Department for Environment, Food and Rural Affairs (Defra) for funding. The authors gratefully acknowledge EMEP for the availability of the ozone datasets.

References

AQEG, 2009. Ozone in the United Kingdom: Air Quality Expert Group, Defra Publications, London. http://www.defra.gov.uk/environment/quality/air/airquality/publications/ozone/documents/aqeg-ozone-report.pdf.

Austin, E., Coull, B., Zanobetti, A., Koutrakis, P., 2013. A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. Environ. Int. 59, 244-254.

Carslaw, D. C., Ropkins, K., 2013. openair: Open-source tools for the analysis of air pollution data. R package version 0.8-5.

Chan, E., Vet, R. J., 2010. Baseline levels and trends of ground level ozone in Canada and the United States. Atmos. Chem. Phys. 10, 8629-8647.

Dabboor, M., Yackel, J., Hossain, M., Braun, A., 2013. Comparing matrix distance measures for unsupervised POLSAR data classification of sea ice based on agglomerative clustering. Int. J. Remote Sens. 34, 1492-1505.

Dillner, A. M., Schauer, J. J., Christensen, W. F., Cass, G. R., 2005. A quantitative method for clustering size distributions of elements. Atmos. Environ. 39, 1525-1537.

EMEP/CCC-Report 1/95, Revision 1/2002. 2002. EMEP manual for sampling and chemical analysis, NILU, Norway, http://www.nilu.no/projects/ccc/manual/.

Flemming, J., Stern, R., Yamartino, R. J., 2005. A new air quality regime classification scheme for O-3, NO2, SO2 and PM10 observations sites. Atmos. Environ. 39, 6121-6129.

Gerasopoulos, E., Kouvarakis, G., Vrekoussis, M., Kanakidou, M., Mihalopoulos, N., 2005. Ozone variability in the marine boundary layer of the eastern Mediterranean based on 7-year observations. J. Geophys. Res. 110, D15309, doi:10.1029/2005JD005991.

Gerasopoulos, E., Kouvarakis, G., Vrekoussis, M., Donoussis, C., Mihalopoulos, N., Kanakidou, M., 2006. Photochemical ozone production in the eastern Mediterranean. Atmos. Environ. 40, 3057-3069.

Guo, H., Ling, H., Cheung, K., Jiang, F., Wang, D. W., Simpson, I. J., Barletta, B., Meinardi, S., Wang, T. J., Wang, X. M., Saunders, S. M., Blake, D. R., 2013. Characterization of photochemical pollution at different elevations in mountainous areas in Hong Kong. Atmos. Chem. Phys. 13, 3881 - 3898.

Henne, S., Brunner, D., Folini, D., Solberg, S., Klausen, J., Buchmann, B., 2010. Assessment of parameters describing representativeness of air quality in-situ measurement sites. Atmos. Chem. Phys. 10, 3561-3581.

Ignaccolo, R., Ghigo, S., Giovenali, E., 2008. Analysis of air quality monitoring networks by functional clustering. Environmetrics 19, 672-686.

Jenkin, M. E., 2008. Trends in ozone concentration distributions in the UK since 1990: Local, regional and global influences. Atmos. Environ. 42, 5434-5445.

Joly, M., Peuch, V. H., 2012. Objective classification of air quality monitoring sites over Europe. Atmos. Environ. 47, 111-123.

Kaufman, L., Rousseeuw, P. J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Kovac-Andric, E., Sorgo, G., Kezele, N., Cvitas, T., Klasinc, L., 2010. Photochemical pollution indicators-an analysis of 12 European monitoring stations. Environ. Monit. Assess. 165, 577-583.

Lau, J., Hung, W. T., Cheung, C. S., 2009. Interpretation of air quality in relation to monitoring station's surroundings. Atmos. Environ. 43, 769–777.

Lee, D. D., Seung, H. S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788-791.

Lee, D. D., Seung, H. S., 2001. Algorithms for non-negative matrix factorization. Adv. Neural Inf. Process. Syst. 13, 556-562.

Lee, J. D., Lewis, A. C., Monks, P. S., Jacob, M., Hamilton, J. F., Hopkins, J. R., Watson, N. M., Saxton, J. E., Ennis, C., Carpenter, L. J., Carslaw, N., Fleming, Z., Bandy, B. J., Oram, D. E., Penkett, S. A., Slemr, J., Norton, E., Rickard, A. R., Whalley, L. K., Heard, D. E., Bloss, W.

J., Gravestock, T., Smith, S. C., Stanton, J., Pilling, M. J., Jenkin, M. E., 2006. Ozone photochemistry and elevated isoprene during the UK heatwave of August 2003. Atmos. Environ. 40, 7598-7613.

Liu, S., 2012. NMFN: Non-negative Matrix Factorization. R package version 2.0. http://CRAN.R-project.org/package=NMFN.

Lu, H. C., Chang, C. L., Hsieh, J. C., 2006. Classification of PM10 distributions in Taiwan. Atmos. Environ. 40, 1452-1463.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2013. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.

Mangiameli, P., Chen, S. K., West, D., 1996. A comparison of SOM neural network and hierarchical clustering methods. Eur. J. Oper. Res. 93, 402-417.

McFadyen, G. G., Cape, J. N., 2005. Peroxyacetyl nitrate in eastern Scotland. Sci. Total Environ. 337, 213-222.

Parrish, D. D., Law, K. S., Staehelin, J., Derwent, R., Cooper, O. R., Tanimoto, H., Volz-Thomas, A., Gilge, S., Scheel, H. E., Steinbacher, M., Chan, E., 2013. Lower tropospheric ozone at northern midlatitudes: Changing seasonal cycle. Geophys. Res. Lett. 40, 1631-1636.

R Core Development Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. .

RoTAP, 2012. Review of Transboundary Air pollution: Acidification, Eutrophication, Ground Level Ozone and Heavy metals in the UK. Contract Report to the Department for Environment, Food and Rural Affairs. Centre for Ecology and Hydrology. http://www.rotap.ceh.ac.uk/sites/rotap.ceh.ac.uk/files/CEH%20RoTAP.pdf.

Royal Society, 2008. Ground-level ozone in the 21st century: future trends, impacts and policy implications. The Royal Society, London. (Science Policy, 15/08). http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/publications/2008/7925.pdf.

Spangl, W., Schneider, J., Moosmann, L., Nagi, C., 2007. Representativeness and Classification of Air Quality Monitoring Stations, Umweltbundesamt Report. http://www.umweltbundesamt.at/fileadmin/site/publikationen/REP0121.pdf.

Tarasova, O. A., Brenninkmeijer, C. A. M., Joeckel, P., Zvyagintsev, A. M., Kuznetsov, G. I., 2007. A climatology of surface ozone in the extra tropics: cluster analysis of observations and model results. Atmos. Chem. Phys. 7, 6099-6117.

Torseth, K., Aas, W., Breivik, K., Fjaeraa, A. M., Fiebig, M., Hjellbrekke, A. G., Myhre, C. L., Solberg, S., Yttri, K. E., 2012. Introduction to the European Monitoring and Evaluation Programme (EMEP) and observed atmospheric composition change during 1972-2009. Atmos. Chem. Phys. 12, 5447-5481.

Ward, J., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236 - 244.

Wilson, R. C., Fleming, Z. L., Monks, P. S., Clain, G., Henne, S., Konovalov, I. B., Szopa, S., Menut, L., 2012. Have primary emission reduction measures reduced ozone across Europe? An analysis of European rural background ozone trends 1996-2005. Atmos. Chem. Phys. 12, 437-454.

Table 1: Number of sites used in cluster analysis for each four year period. The increasing number of countries with sites indicates the increasing geographical coverage across Europe with time. 49 sites are common to all time periods.

Time period	**No. of sites**	**No. of countries**
1991-1994	76	14
1995-1998	100	20
1999-2002	117	27
2003-2006	117	27
2007-2010	113	26

Figure 1: Illustration of the process of non-negative factorization as applied to the ozone data in this work.

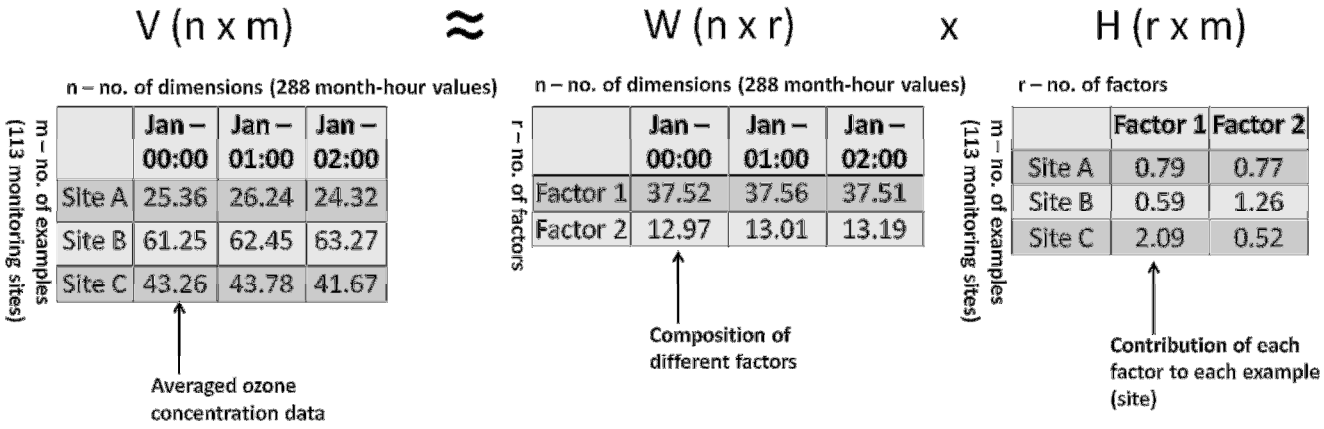


Figure 2: Average ozone monthly-diurnal cycles of two factors produced during non-negative matrix factorisation of data for 2007 – 2010. Concentrations are µg m^{-3}.

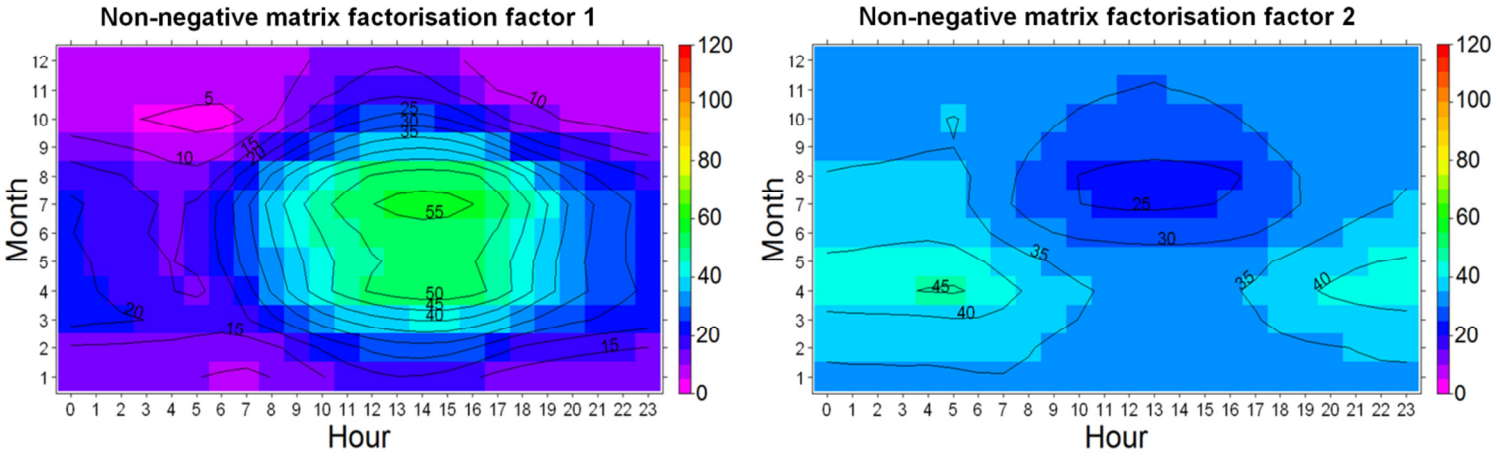


Figure 3: Dendrogram of 2007-10 EMEP sites derived by Ward's method of hierarchical clustering and reordered using non-negative matrix factorisation with the two factors whose monthly-diurnal ozone concentrations are illustrated in Figure 2. The UK EMEP sites are identified with a red dot for those classified as Remote and a green dot for those classified as Polluted. The two UK EMEP sites of Harwell and Auchencorth are circled.

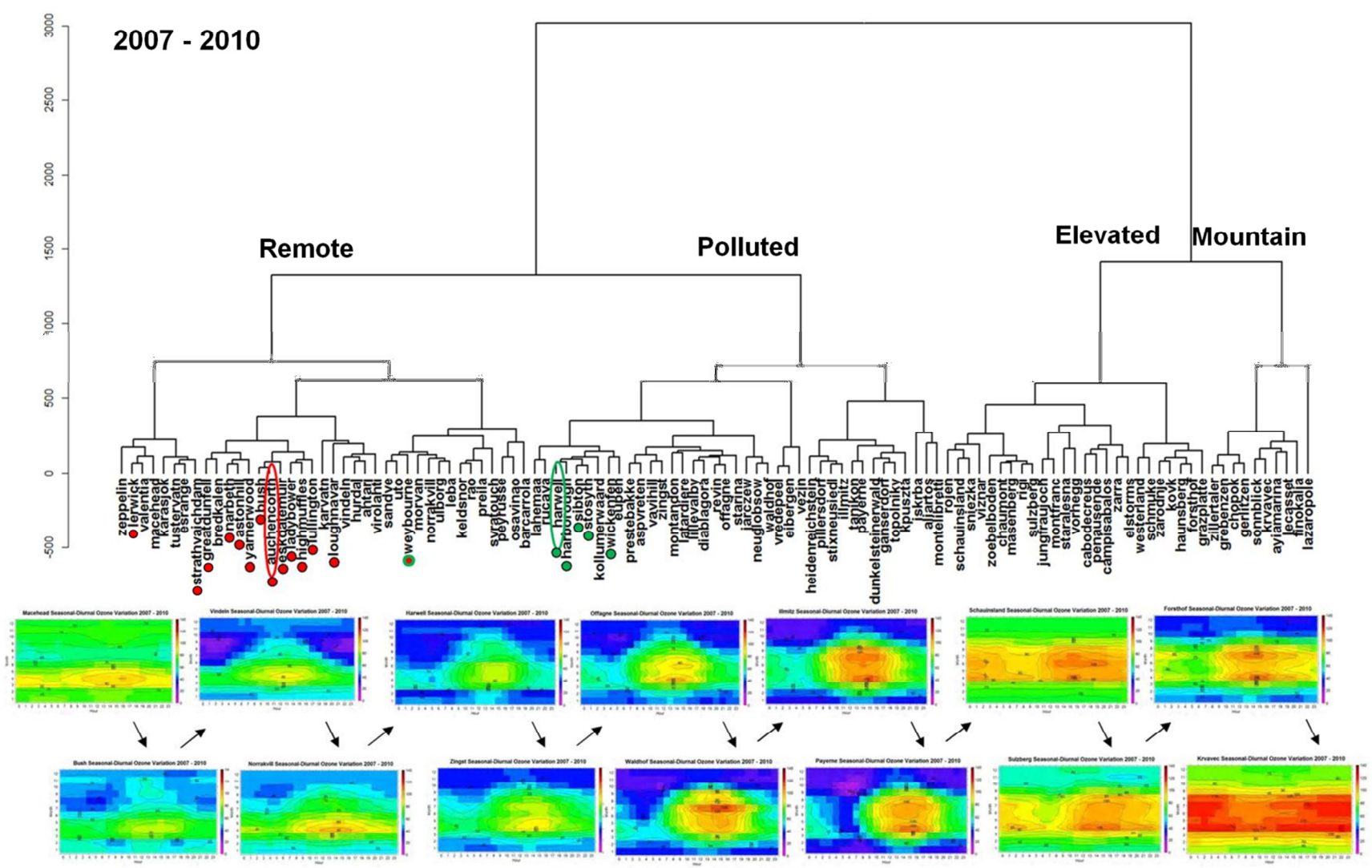


Figure 4: The proportion of within-cluster variance explained as a function of number of clusters (2007-2010 dataset).

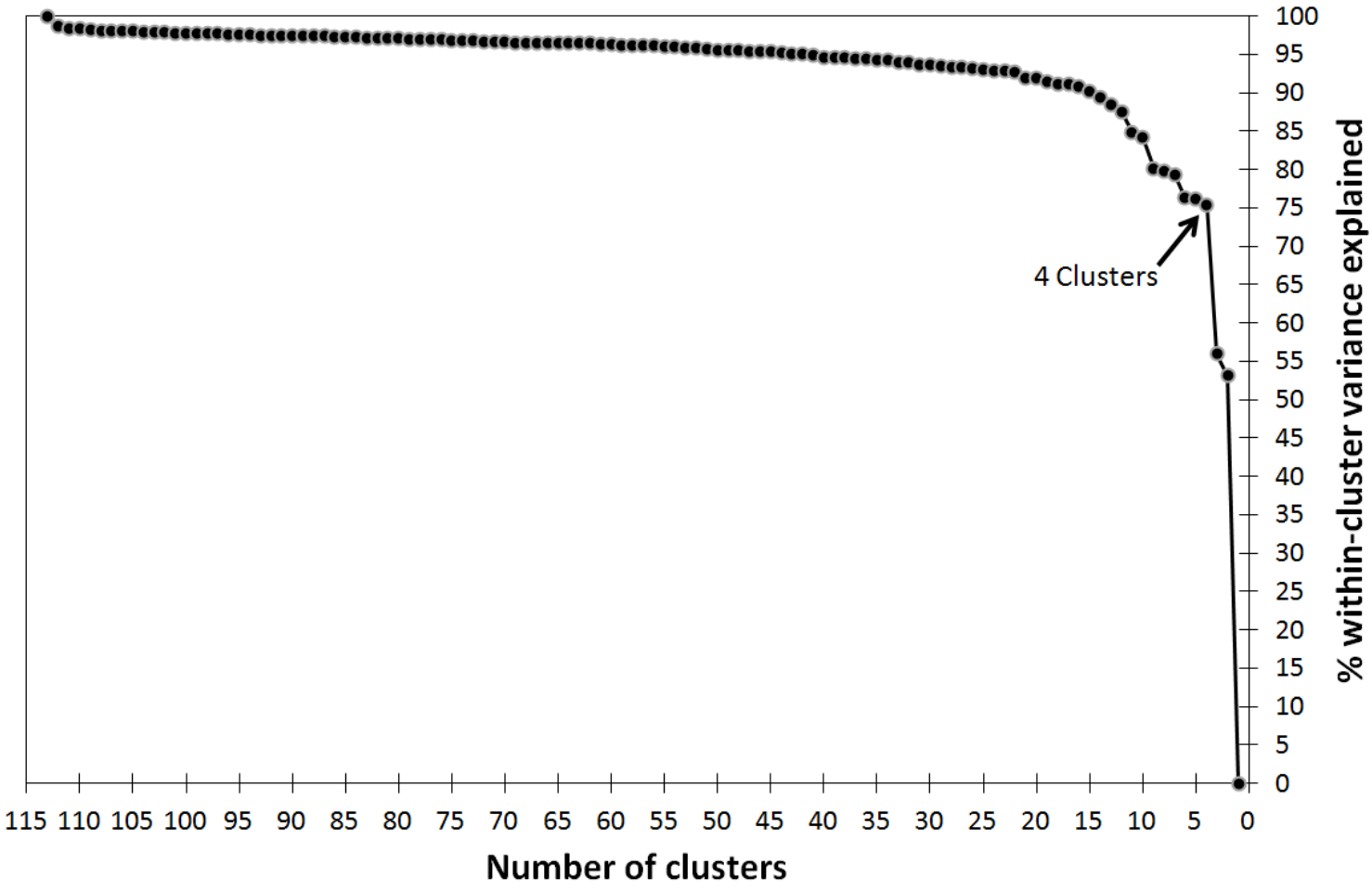


Figure 5: Average ozone monthly-diurnal cycle for the four clusters assigned for 2007 – 2010. Concentrations are µg m^{-3}.

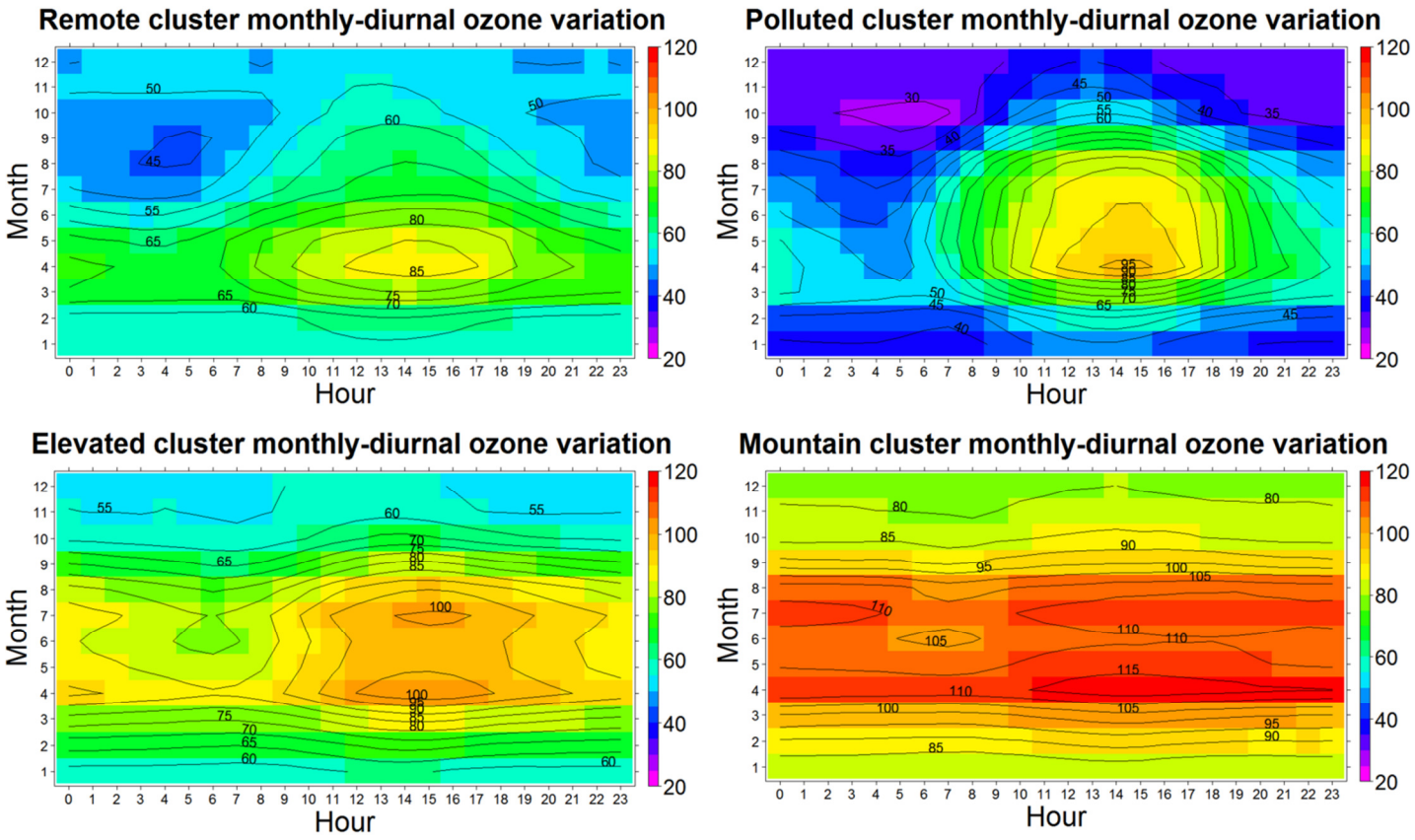
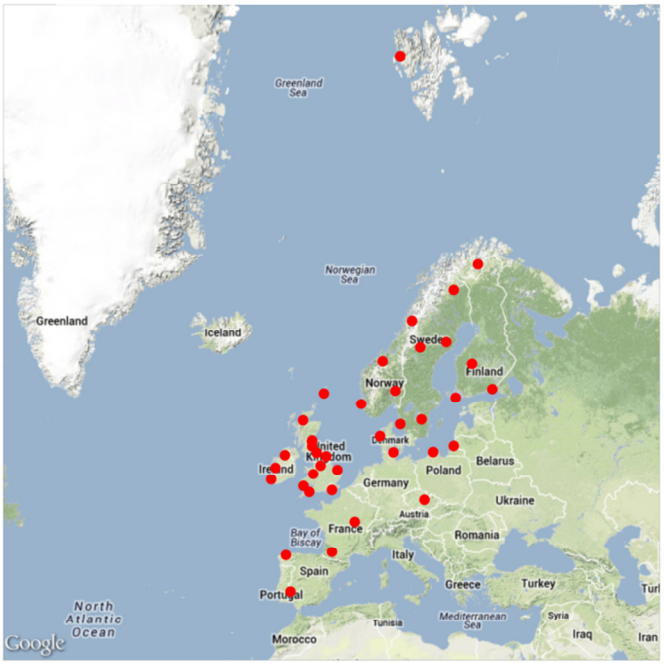


Figure 6: Locations of the 113 EMEP sites separated according to the four clusters assigned for 2007 – 2010 monthly-diurnal ozone cycles (Map data: Google, Basarsoft, GeoBasis-DE/BKG, ORION-ME).

2007 – 2010 Remote cluster site locations



2007 – 2010 Polluted cluster site locations

2007 – 2010 Elevated cluster site locations

2007 – 2010 Mountain cluster site locations

Figure 7: Location of UK EMEP sites operational for 2007 – 2010. Sites clustered as Remote are shown in red, and those clustered as Polluted are shown in green (Map data: Google, GeoBasis-DE/BKG).